Submitted to the Annals of Applied Probability

NONPARAMETRIC ESTIMATION BY CONVEX PROGRAMMING

By Anatoli B. Juditsky and Arkadi S.Nemirovski*

Université Grenoble I and Georgia Institute of Technology

The problem we concentrate on is as follows: given 1) a convex compact set X in $\mathbb{R}^n$, an affine mapping $x \mapsto A(x)$, a parametric family $\{p_\mu(\cdot)\}$ of probability densities; 2) N i.i.d. observations of the random variable ω, distributed with the density $p_{A(x)}(\cdot)$ for some (unknown) $x \in X$; estimate the value $g^T x$ of a given linear form at x.

For several families $\{p_\mu(\cdot)\}$ with no additional assumptions on X and A, we develop computationally efficient estimation routines which are minimax optimal, within an absolute constant factor. We then apply these routines to recovering x itself in the Euclidean norm.

1. Introduction. The problem we are interested in is essentially as follows: suppose that we are given a convex compact set X in $\mathbb{R}^n$, an affine mapping $x \mapsto A(x)$ and a parametric family $\{p_\mu(\cdot)\}$ of probability densities. Suppose that N i.i.d. observations of the random variable ω, distributed with the density $p_{A(x)}(\cdot)$ for some (unknown) $x \in X$, are available. Our objective is to estimate the value $g^T x$ of a given linear form at x.

In Nonparametric Statistics, there exists an immense literature on various versions of this problem, see, e.g. [8, 10, 11, 13, 16, 17, 26–31, 33] and the references therein. To the best of our knowledge, the majority of papers on the subject focus on specific domains X (e.g., distributions with densities from Sobolev balls), and investigate lower and upper bounds on the worst-case, w.r.t. $x \in X$, accuracy to which the problem of interest can be solved. These bounds depend on the number of observations N, and the question of primary interest is the behavior of those bounds as $N \to \infty$. When the lower and the upper bounds coincide within a constant factor (or, ideally, within factor $(1+o(1))$) as $N \to \infty$), the estimation problem is considered as being essentially solved, and the estimation methods underlying the upper bounds are treated as optimal.

* Research of the second author was partly supported by the NSF grant DMI # 0619977

AMS 2000 subject classifications: Primary 62G08; secondary 62G15, 62G07

Keywords and phrases: estimation of linear functional, minimax estimation, oracle inequalities, convex optimization, PE tomography

The approach we adopt in this paper is of a different spirit: we make no "structural assumptions" on X, aside of crucial for us assumptions of convexity and compactness, and we make no assumptions on the linear functional p. Clearly, with no structural assumptions on X and p, explicit bounds on the risks of our estimates, same as bounds on the minimax optimal risk, are impossible. However, it is possible to show that *when estimating linear forms, the worst-case risk of the estimator we propose is within an absolute constant factor of the "ideal"* (i.e., the minimax optimal) *risk*. It should be added that while the optimal, within an absolute constant factor, worst-case risk of our estimates is not available in a closed analytical form, it is "available algorithmically" – it can be efficiently computed, provided that X is computationally tractable[1].

Note that the estimation problem, presented above, can be seen as a generalization of the problem of estimation of linear functionals of central parameter of normal distribution [14]. Namely, suppose that the observation $\omega \in \mathbb{R}^m$,

$$\omega = Ax + \sigma\xi$$

of unknown signal x is available. Here A is a given $m \times n$ matrix and $\xi \sim \mathcal{N}(0, I_m)$, $\sigma > 0$ is known. For this important case the problem has been essentially solved in [7], where it was proved that for several commonly used loss functions, the minimax optimal *affine in* ω estimate is minimax optimal, within an absolute constant factor, among *all possible* estimates.

Another special case of our setting is the problem of estimating a linear functional $g(p)$ of an unknown distribution p, given N i.i.d observations $\omega_1, ..., \omega_N$ which obey p. We suppose that it is known *a priori* that $p \in X$, where X is a given convex compact set of distributions (here the parameter x is the density p itself). Some important results for this problem has been obtained in [4] and [5]. For instance, in [5] the authors established minimax bounds for the risk of estimation of $g(p)$ and developed the estimation method based on the binary search algorithm. The estimation procedure uses at each search iteration tests of convex hypotheses, studied in [2, 3]. That estimator of $g(p)$ is shown to be minimax optimal (within an absolute constant factor) if some basic structural assumptions about X hold.

In this paper we concentrate on the properties of *affine estimators*.[2] Our motivation is to extend the results, obtained in [7] for Gaussian case. In

[1] For details on computational tractability and complexity issues in Convex Optimization, see, e.g., [1, Chapter 4]. A reader not familiar with this area will not lose much when interpreting a computationally tractable convex set as a set given by a finite system of inequalities $p_i(x) \leq 0$, $i = 1, ..., m$, where $p_i(x)$ are convex polynomials.

[2] We refer to an estimator as affine if it is affine function of empirical distribution.

particular, we propose a technique of derivation of affine estimators which are minimax optimal (up to an absolute constant) for the class of a "good parametric family of distribution", which is defined in Section 2.1. As normal family and discrete distributions belong to the class of good parametric families, the minimax optimal estimators for these cases are obtained by direct application of general construction. In this sense, our results generalize those of [5] and [7] on estimation of linear functionals. On the other hand, it is clear that different techniques, presented in the current paper, inherit from those developed in [3] and [5]. To make a computationally efficient solution of the estimation problem possible, unlike the authors of those papers, we concentrate only on the finite-dimensional situation. As a result, the proposed estimation procedures allows efficient numeric implementation. However, we allow the dimension to be arbitrarily large, thus addressing, essentially, a nonparametric estimation problem.

The rest of this paper is organized as follows. In Section 2 we define the main components of our study – we state the estimation problem and define the corresponding risk measures. Then Section 3 we provide the general solution to the estimation problem which then is applied in Section 4 to different estimation problems: that of estimating linear functionals in normal model and tomography model. In the concluding Sections 5 and 6 we consider adaptive versions of affine estimators and also discuss briefly how our results can be used in order to recover the "entire" signal x underlying our observations.

Note that when passing from recovering linear forms of the unknown signal to recovering the signal itself, we do impose structural assumptions on X, but still make no structural assumptions on the affine mapping $A(x)$. Our "optimality results" become weaker – instead of "optimality within an absolute constant factor" we end up with statements like "the worst-case risk of such-and-such estimate is in-between the minimax optimal risk and the latter risk to the power χ", with χ depending on the geometry of X (and close to 1 when this geometry is "good enough").

2. Problem statement.

2.1. *Good parametric families of distributions.* Let (Ω, P) be a Polish space with Borel σ-finite measure, and $\mathcal{M} \subset \mathbb{R}^m$. Assume that every $\mu \in \mathcal{M}$ is associated with a probability density $p_\mu(\omega)$ – a Borel nonnegative function on Ω such that $\int_\Omega p_\mu(\omega)P(d\omega) = 1$; we refer to the mapping $\mu \to p_\mu(\cdot)$ as to a *parametric density family* $\mathcal{D}$. Let also $\mathcal{F}$ be a finite-dimensional linear space of Borel functions on Ω which contains constants. We call a pair $(\mathcal{D}, \mathcal{F})$ *good*, if it possesses the following properties:

1. $\mathcal{M}$ is an open convex set in $\mathbb{R}^m$;
2. Whenever $\mu \in \mathcal{M}$, we have $p_\mu(\omega) > 0$ everywhere on Ω
3. Whenever $\mu, \nu \in \mathcal{M}$, we have $\phi(\omega) = \ln(p_\mu(\omega)/p_\nu(\omega)) \in \mathcal{F}$
4. Whenever $\phi(\omega) \in \mathcal{F}$, the function

$$F_\phi(\mu) = \ln\left(\int_\Omega \exp\{\phi(\omega)\} p_\mu(\omega) P(d\omega)\right)$$

is well defined and concave in $\mu \in \mathcal{M}$.

Let us list several examples.

Example 1: Discrete distributions. Let $\Omega = \{1, 2, ..., M\}$ be a finite set, P be a counting measure on Ω, $\mathcal{M} = \{\mu \in \mathbb{R}^M : \mu > 0, \sum_i \mu_i = 1\}$, and $p_\mu(i) = \mu_i$, $i = 1, ..., M$. Let also $\mathcal{F}$ be the set of all functions on Ω. The associated pair $(\mathcal{D}, \mathcal{F})$ clearly is good.

Example 2: Poisson distributions. Let $\Omega = \{0, 1, ...\}$, P be the counting measure on Ω, $\mathcal{M} = \{\mu \in \mathbb{R} : \mu > 0\}$ and $p_\mu(i) = \frac{\mu^i \exp\{-\mu\}}{i!}$, $i \in \Omega$, so that p_μ is the Poisson distribution with the parameter μ. Let also $\mathcal{F}$ be the set of affine functions $\phi(i) = \alpha i + \beta$ on Ω. We claim that the associated pair $(\mathcal{D}, \mathcal{F})$ is good. Indeed, $\ln(p_\mu(i)/p_\nu(i)) = i[\ln \mu - \ln \nu] + \mu - \nu$ is an affine function of i, and

$$\begin{aligned} \ln\left(\sum_i \exp\{\alpha i + \beta\} \frac{\mu^i \exp\{-\mu\}}{i!}\right) &= \ln\left(\exp\{\beta - \mu\} \exp\{\mu \exp\{\alpha\}\}\right) \\ &= \beta - \mu + \mu \exp\{\alpha\} \end{aligned}$$

is a concave function of $\mu > 0$.

Example 3: Gaussian distributions with fixed covariance. Let $\Omega = \mathbb{R}^k$, P be the Lebesque measure on Ω, let Σ be a positive definite $k \times k$ matrix, let $\mathcal{M} = \mathbb{R}^k$ and

$$p_\mu(\omega) = (2\pi)^{-k/2} (\mathrm{Det}\Sigma)^{-1/2} \exp\{-(\omega - \mu)^T \Sigma^{-1} (\omega - \mu)\}$$

be the density of the Gaussian distribution with mean μ and covariance matrix Σ. Let, further, $\mathcal{F}$ be comprised of affine functions on Ω. We claim that the associated pair $(\mathcal{D}, \mathcal{F})$ is good. Indeed, the function $\ln(p_\mu(\omega)/p_\nu(\omega))$ indeed is affine on Ω, and

$$\ln\left(\int \exp\{\phi^T \omega + c\} p_\mu(\omega) d\omega\right) = c + \phi^T \mu + \frac{\phi^T \Sigma \phi}{2}$$

is a concave function of μ.

Example 4: Direct product of good pairs. Let $p^\ell_{\mu_\ell}(\omega_\ell)$ be a probability density, parameterized by $\mu_\ell \in \mathcal{M}_\ell \subset \mathbb{R}^{m_\ell}$, on a Polish space Ω_ℓ with Borel σ-finite measure P_ℓ, and $\mathcal{F}_\ell$ be a finite-dimensional linear space of Borel functions on Ω_ℓ such that the associated pairs $(\mathcal{D}_\ell, \mathcal{F}_\ell)$ are good. Let us define the *direct product* $(\mathcal{D}, \mathcal{F}) = \bigotimes_{\ell=1}^{L} (\mathcal{D}_\ell, \mathcal{F}_\ell)$ of these pairs as follows:

- The associated space with measure is $(\Omega = \Omega_1 \times ... \times \Omega_L, P = P_1 \times ... \times P_\ell)$;
- The set of parameters is $\mathcal{M} = \mathcal{M}_1 \times ... \times \mathcal{M}_L$, and the density associated with a parameter $\mu = (\mu_1, ..., \mu_L)$ from this set is $p_\mu(\omega_1, ..., \omega_L) = \prod_{\ell=1}^L p^\ell_{\mu_\ell}(\omega_\ell)$;
- $\mathcal{F}$ is comprised of all functions $\phi(\omega_1, ..., \omega_L) = \sum_{\ell=1}^L \phi_\ell(\omega_\ell)$ with $\phi_\ell(\cdot) \in \mathcal{F}_\ell$, $\ell = 1, ..., m$.

We claim that *the direct product of good pairs is good*. Indeed, $\mathcal{M}$ is an open convex set; when $\mu = (\mu_1, ..., \mu_L)$ and $\nu = (\nu_1, ..., \nu_L)$ are in $\mathcal{M}$, we have

$$\ln\left(p_\mu(\omega_1, ..., \omega_L)/p_\nu(\omega_1, ..., \omega_L)\right) = \sum_{\ell=1}^{L} \ln(p^\ell_{\mu_\ell}(\omega_\ell)/p^\ell_{\nu_\ell}(\omega_\ell)) \in \mathcal{F},$$

and when $\phi(\omega_1, ..., \omega_L) = \sum_\ell \phi_\ell(\omega_\ell) \in \mathcal{F}$, we have

$$\begin{array}{l}\ln\left(\int_\Omega \exp\{\phi(\omega)\} p_\mu(\omega) P(d\omega)\right) = \ln\left(\prod_\ell \int_{\Omega_\ell} \phi_\ell(\omega_\ell) p^\ell_{\mu_\ell}(\omega_\ell) P(d\omega_\ell)\right) \\ = \sum_\ell \ln\left(\int_{\Omega_\ell} \phi_\ell(\omega_\ell) p^\ell_{\mu_\ell}(\omega_\ell) P(d\omega_\ell)\right)\end{array}$$

is a sum of concave functions of μ and thus is concave in μ.

2.2. *The problem.* The problem we are interested in is as follows:

Problem I. *We are given*

- *a convex compact set $X \subset \mathbb{R}^n$,*
- *a good pair $(\mathcal{D}, \mathcal{F})$ comprised of*
  - *a parametric family $\{p_\mu(\omega) : \mu \in \mathcal{M} \subset \mathbb{R}^m\}$ of probability densities on a Borel space Ω with σ-finite Borel measure P and*
  - *a finite-dimensional linear space $\mathcal{F}$ of Borel functions on Ω*
- *an affine mapping $x \mapsto A(x) : X \mapsto \mathcal{M}$*
- *a linear form $g^T z$ on $\mathbb{R}^n \supset X$.*

Aside of this a priori information, we are given a realization ω of a random variable taking values in Ω and distributed with the density $p_{A(x)}(\cdot)$ for some unknown in advance $x \in X$. Our goal is to infer from this observation an estimate $\widehat{g}(\omega)$ of the value $g^T x$ of the given linear form at x.

We call an estimate *affine*, if it is of the form $\widehat{g}(\omega) = \phi(\omega)$ with certain $\phi \in \mathcal{F}$.

We quantify the risk of a candidate estimate $\widehat{g}(\cdot)$ by its worst-case, over $x \in X$, confidence interval, given the confidence level. Specifically, given a *confidence level* $\epsilon \in (0,1)$, we define the associated *ϵ-risk* of an estimate $\widehat{g}$ as

$$\mathrm{Risk}(\widehat{g};\epsilon) = \inf\left\{\delta : \sup_{x\in X} \mathrm{Prob}_{\omega\sim p_{A(x)}(\cdot)}\{\omega : |\widehat{g}(\omega) - g^T x| \le \delta\} < \epsilon\right\}. \tag{2.1}$$

The corresponding minimax optimal ϵ-risk is defined as

$$\mathrm{Risk}_*(\epsilon) = \inf_{\widehat{g}(\cdot)} \mathrm{Risk}(\widehat{g};\epsilon), \tag{2.2}$$

where inf is taken over the space of all Borel functions $\widehat{g}$ on Ω. We are interested also in the minimax optimal ϵ-risk of affine estimates

$$\mathrm{RiskA}(\epsilon) = \inf_{\phi(\cdot)\in\mathcal{F}} \mathrm{Risk}(\phi;\epsilon). \tag{2.3}$$

3. Minimax optimal affine estimators.

3.1. *Main result.* Our main result is as follows:

THEOREM 3.1. *Let the pair $(\mathcal{D},\mathcal{F})$ underlying Problem I be good. Then the minimax optimal risk achievable with affine estimates is, for small ϵ, within an absolute constant factor of the "true" minimax optimal risk, specifically,*

$$0 \le \epsilon < 1/4 \Rightarrow \mathrm{RiskA}(\epsilon) \le \theta(\epsilon)\,\mathrm{Risk}_*(\epsilon),\ \theta(\epsilon) = \frac{2\ln\left(\frac{2}{\epsilon}\right)}{\ln\left(\frac{1}{4\epsilon}\right)}. \tag{3.1}$$

PROOF. For $r \ge 0$, let us set

(3.2)

$$\begin{array}{rcl}
\Phi_r(x,y;\phi,\alpha) & = & g^T x - g^T y + \alpha\ln\left(\int_\Omega \exp\{\alpha^{-1}\phi(\omega)\}p_{A(y)}(\omega)P(d\omega)\right)\\
& & +\alpha\ln\left(\int_\Omega \exp\{-\alpha^{-1}\phi(\omega)\}p_{A(x)}(\omega)P(d\omega)\right) + 2\alpha r : Z\times\mathcal{F}_+ \to \mathbb{R},\\
Z & = & X\times X,\\
\mathcal{F}_+ & = & \mathcal{F}\times\{\alpha>0\}.
\end{array}$$

We claim that this function is a continuous real-valued function on $Z\times\mathcal{F}_+$ which is convex in $(\phi,\alpha)\in\mathcal{F}_+$ and concave in $(x,y)\in Z$.

Indeed, the function

$$\begin{array}{rcl}\Psi(\mu,\nu;\phi) & = & \ln\left(\int_\Omega \exp\{\phi(\omega)\}p_\mu(\omega)P(d\omega)\right)\\ & & +\ln\left(\int_\Omega \exp\{-\phi(\omega)\}p_\nu(\omega)P(d\omega)\right) : (\mathcal{M}\times\mathcal{M})\times\mathcal{F}\to\mathbb{R}\end{array}$$

is well-defined, concave in $(\mu,\nu)\in\mathcal{M}\times\mathcal{M}$ (since $(\mathcal{D},\mathcal{F})$ is good) and convex in $\phi\in\mathcal{F}$ (evident). Since $\mathcal{M}$ is open and $\mathcal{F}$ is a finite-dimensional linear space, Ψ is continuous on its domain. It remains to note that Φ_ϵ is the sum of a linear function of x,y,α and the function $\alpha\Psi(A(x),A(y);\alpha^{-1}\phi)$ which clearly is concave in (x,y) (since $\Psi(\mu,\nu;\phi)$ is concave in (μ,ν) and $A(\cdot)$ is affine) and convex in $(\phi,\alpha)\in\mathcal{F}_+$ (since $\Psi(\mu,\nu;\phi)$ is continuous in $\phi\in\mathcal{F}$, and the transformation $f(u)\mapsto g(u,\alpha)=\alpha f(u/\alpha)$ converts a convex function of u into a convex in $(\alpha>0,u)$ function of (u,α)).

Since Z is a convex finite-dimensional compact set, $\mathcal{F}_+$ is a convex finite-dimensional set and Φ_ϵ is continuous and convex-concave on $Z\times\mathcal{F}_+$, we can invoke the Sion-Kakutani Theorem (see, e.g., [12]) to infer that

$$\sup_{x,y\in X}\inf_{\phi\in\mathcal{F},\alpha>0}\Phi_r(x,y;\phi,\alpha)=\inf_{\phi\in\mathcal{F},\alpha>0}\max_{x,y\in X}\Phi_r(x,y;\phi,\alpha):=2\Phi_*(r). \tag{3.3}$$

Note that $\Phi_*(r)\geq 0$ is a concave and nonnegative function of $r\geq 0$. Indeed, the functional $f_x[h]=\ln\int_\Omega\exp\{h(\omega)\}p_{A(x)}(\omega)P(d\omega)$ is well defined and convex on $\mathcal{F}$, whence

$$\Phi_r(x,x;\phi,\alpha)=2\alpha r+\alpha\left(f_x[-\alpha^{-1}\phi]+f_x[\alpha^{-1}\phi]\right)\geq 2\alpha r\geq 0,$$

whence $\Phi_*(r)\geq\frac{1}{2}\sup_{x\in X}\inf_{\phi\in\mathcal{F},\alpha>0}\Phi_r(x,x;\phi,\alpha)\geq 0$. The concavity of $\Phi_*(r)$ on the nonnegative ray follows immediately from the representation, yielded by (3.3),

$$\Phi_*(r)=\frac{1}{2}\inf_{\phi\in\mathcal{F},\alpha}\left[2\alpha r+\sup_{x,y\in X}\Phi_0(x,y;\phi,\alpha)\right]$$

of $\Phi_*(r)$ as the infinum of a family of affine functions of r.

LEMMA 3.1. *One has*

$$\mathrm{RiskA}(\epsilon)\leq\Phi_*(\ln(2/\epsilon)). \tag{3.4}$$

PROOF. Given $\delta>0$ and $\epsilon\in(0,1/4)$, let us build an affine estimate with ϵ-risk $\leq R\equiv\Phi_*(\ln(2/\epsilon))+\delta/2$, namely, as follows. By (3.3), there exist

$\phi_* \in \mathcal{F}$ and $\alpha_* > 0$ such that

$$\begin{array}{l} 2\Phi_*(\ln(2/\epsilon)) + \delta/2 \geq \max_{x,y\in X} \Phi_{\epsilon/2}(x,y;\phi_*,\alpha_*) \\ = \underbrace{\max\limits_{x\in X}\left[g^T x + \alpha_* \ln\left(\int_\Omega \exp\{-\alpha_*^{-1}\phi_*(\omega)\}p_{A(x)}(\omega)P(d\omega)\right) + \alpha_*\ln(2/\epsilon)\right]}_{U} \\ \underbrace{\max\limits_{y\in X}\left[-g^T y + \alpha_* \ln\left(\int_\Omega \exp\{\alpha_*^{-1}\phi_*(\omega)\}p_{A(y)}(\omega)P(d\omega)\right) + \alpha_*\ln(2/\epsilon)\right]}_{V} \end{array}$$

Setting $c = \frac{U-V}{2}$, we have

$$\begin{array}{l} \max_{x\in X}\left[g^T x + \alpha_* \ln\left(\int_\Omega \exp\{-\alpha_*^{-1}[\phi_*(\omega)+c]\}p_{A(x)}(\omega)P(d\omega)\right) + \alpha_*\ln(2/\epsilon)\right] \\ = U - c = \frac{U+V}{2} \leq \Phi_*(\ln(2/\epsilon)) + \delta/4 = R - \delta/4, \\ \max_{y\in Y}\left[g^T x + \alpha_* \ln\left(\int_\Omega \exp\{\alpha_*^{-1}[\phi_*(\omega)+c]\}p_{A(y)}(\omega)P(d\omega)\right) + \alpha_*\ln(2/\epsilon)\right] \\ = V + c = \frac{U+V}{2} \leq \Phi_*(\ln(2/\epsilon)) + \delta/4 = R - \delta/4, \end{array}$$

or, equivalently,

$$\begin{array}{l} \max_{x\in X} \ln\left(\int_\Omega \exp\left\{\alpha_*^{-1}\left[g^T x - (\phi_*(\omega)+c) - R\right]\right\} p_{A(x)}(\omega)P(d\omega)\right) \\ \leq \ln(\epsilon/2) - \frac{\delta}{4\alpha_*} \equiv \ln(\epsilon'/2), \\ \max_{y\in X} \ln\left(\int_\Omega \exp\left\{\alpha_*^{-1}\left[(\phi_*(\omega)+c) - R - g^T y\right]\right\} p_{A(y)}(\omega)P(d\omega)\right) \leq \ln(\epsilon'/2), \end{array}$$

that is,

(3.5)

$$\begin{array}{ll} (a) & \forall x \in X : \int_\Omega \exp\left\{\alpha_*^{-1}\left[g^T x - (\phi_*(\omega)+c) - R\right]\right\} p_{A(x)}(\omega)P(d\omega) \leq \epsilon'/2, \\ (b) & \forall y \in X : \int_\Omega \exp\left\{\alpha_*^{-1}\left[[\phi_*(\omega)+c] - R - g^T y\right]\right\} p_{A(y)}(\omega)P(d\omega) \leq \epsilon'/2. \end{array}$$

For a given $x \in X$, the exponent in (a) is nonnegative and is > 1 for all ω such that $g^T x - [\phi_*(\omega)+c] > R$; therefore (a) implies that $\mathrm{Prob}_{\omega\sim p_{A(x)}(\cdot)}\{g^T x > [\phi_*(\omega)+c] + R\} \leq \epsilon'/2$ for every $x \in X$. By similar reasons, (b) implies that $\mathrm{Prob}_{\omega\sim p_{A(x)}(\cdot)}\{g^T x < [\phi_*(\omega)+c] - R\} \leq \epsilon'/2$ for all $x \in X$. Since by construction $\epsilon' < \epsilon$, we see that the ϵ-risk of the affine estimate $\widehat{g}(\omega) = \phi_*(\omega) + c$ is $\leq R$, as claimed. ∎

LEMMA 3.2. *One has*

$$\delta \in (0,1) \Rightarrow \mathrm{Risk}_*(\delta^2/4) \geq \Phi_*(\ln(1/\delta)), \tag{3.6}$$

whence also

$$\epsilon \in (0,1/4) \Rightarrow \mathrm{Risk}_*(\delta) \leq \frac{2\ln\left(\frac{2}{\epsilon}\right)}{\ln\left(\frac{1}{4\epsilon}\right)}\Phi_*(\ln(2/\epsilon)). \tag{3.7}$$

PROOF. To prove (3.6), let us set $\rho = \ln(1/\delta)$. The function $\Psi_\rho(x,y) = \inf_{\phi\in\mathcal{F},\alpha>0}\Phi_\rho(x,y;\phi,\alpha)$ takes values in $\{-\infty\}\cup\mathbb{R}$, is upper semicontinuous (since Φ_r is continuous) and is not identically $-\infty$ (in fact, it is even ≥ 0 when $y = x$). Thus, Ψ_ρ achieves its maximum on $X\times X$ at certain point $(\bar{x},\bar{y})$, and for any $(\alpha>0,\phi\in\mathcal{F})$:

$$\Phi_\rho(\bar{x},\bar{y};\phi,\alpha) \geq \Psi_\rho(\bar{x},\bar{y}) = \sup_{x,y\in X}\inf_{\phi\in\mathcal{F},\alpha>0}\Phi_\rho(x,y;\phi,\alpha) = 2\Phi_*(\rho), \tag{3.8}$$

where the concluding inequality is given by (3.3). Since $(\mathcal{D},\mathcal{F})$ is a good pair, setting $\mu = A(\bar{x})$, $\nu = A(\bar{y})$ and $\bar{\phi}(\omega) = \frac{1}{2}\ln(p_\mu(\omega)/p_\nu(\omega))$, we get $\bar{\phi}\in\mathcal{F}$, which combines with (3.8) to imply that

$$\begin{aligned}
&\forall(\alpha>0):\\
&2\Phi_*(\rho) \leq g^T\bar{x} - g^T\bar{y} + \alpha\Big[\ln\left(\textstyle\int_\Omega \exp\{-\alpha^{-1}[\alpha\bar{\phi}(\omega)]\}p_\mu(\omega)P(d\omega)\right)\\
&+\ln\left(\textstyle\int_\Omega \exp\{\alpha^{-1}[\alpha\bar{\phi}(\omega)]\}p_\nu(\omega)P(d\omega)\right) + 2\rho\Big]\\
&= g^T\bar{x} - g^T\bar{y} + 2\alpha\left[\ln\left(\textstyle\int_\Omega\sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\right)+\rho\right]
\end{aligned}$$

The resulting inequality holds true for all $\alpha>0$, meaning that

$$\begin{array}{ll}(a) & g^T\bar{x} - g^T\bar{y} \geq 2\Phi_*(\rho) = 2\Phi_*(\ln(1/\delta)),\\ (b) & \int_\Omega\sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega) \geq \exp\{-\rho\} = \delta.\end{array} \tag{3.9}$$

Now assume, in contrast to what should be proved, that $\mathrm{Risk}_*(\delta^2/4) < \Phi_*(\ln(1/\delta))$. Then there exists $R' < \Phi_*(\ln(1/\delta))$, $\delta' < \delta^2/4$ and an estimate $\widehat{g}(\omega)$ such that

$$\mathrm{Prob}_{\omega\sim p_{A(x)}(\cdot)}\left\{|\widehat{g}(\omega) - g^Tx| > R'\right\} \leq \delta'\ \forall x\in X.$$

Now consider two hypotheses $\Pi_{1,2}$ on the distribution of ω stating that the densities of the distribution w.r.t. P are p_μ, p_ν, respectively. Consider a procedure for distinguishing between the hypotheses as follows: after ω is observed, we compare $\widehat{g}(\omega)$ with $\bar{g} = \frac{1}{2}[g^T\bar{x}+g^T\bar{y}]$; if $\widehat{g}(\omega)\geq\bar{g}$, we accept Π_1, otherwise we accept Π_2. Note that by (3.9.a) and due to $R' < \Phi_*(\ln(1/\delta))$, the probability to accept Π_2 when Π_1 is true is $\leq$ the probability for $\widehat{g}(\omega)$ to deviate from $g^T\bar{x}$ by at most R', that is, it is $\leq\delta'$. Similarly, the probability to accept Π_1 when Π_2 is true is $\leq\delta'$. Now let Ω_1 be the part of Ω where our hypotheses testing routine accepts Π_1, so that in $\Omega_2\Omega\backslash\Omega_1$ the routine accepts Π_2. As we just have seen,

$$\int_{\Omega_1} p_\nu(\omega)P(d\omega) \leq \delta', \int_{\Omega_2} p_\mu(\omega)P(d\omega) \leq \delta',$$

whence

$$\int_\Omega \sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega) = \sum_{i=1}^2 \int_{\Omega_i} \sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)$$
$$\le \sum_{i=1}^2 \left(\int_{\Omega_i} p_\mu(\omega)P(d\omega)\right)^{1/2}\left(\int_{\Omega_i} p_\nu(\omega)P(d\omega)\right)^{1/2} \le 2\sqrt{\delta'} < 2\sqrt{\delta^2/4} = \delta.$$

The resulting inequality $\int_\Omega \sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega) < \delta$ contradicts (3.9.*b*); we have arrived at a desired contradiction. (3.6) is proved.

To prove (3.7), let us set $\delta = 2\sqrt{\epsilon}$, so that $\mathrm{Risk}_*(\epsilon) = \mathrm{Risk}_*(\delta^2/4) \ge \Phi_*(\ln(1/\delta)) = \Phi_*\left(\frac{1}{2}\ln\left(\frac{1}{4\epsilon}\right)\right)$, where the concluding $\ge$ is due to (3.6). Now recall that $\Phi_*(r)$ is a nonnegative and concave function of $r \ge 0$, so that $\Phi_*(tr) \le t\Phi_*(r)$ for all $r \ge 0$ and $t \ge 1$. We therefore have $\Phi_*\left(\frac{1}{2}\ln\left(\frac{1}{4\epsilon}\right)\right) \ge \frac{\ln\left(\frac{1}{4\epsilon}\right)}{2\ln\left(\frac{2}{\epsilon}\right)}\Phi_*\left(\frac{2}{\epsilon}\right)$, and we arrive at (3.7). ∎

Lemmas 3.1, 3.2 clearly imply Theorem 3.1. ∎

REMARK 3.1. *Lemmas 3.1, 3.2 provide certain information even beyond the case when the pair $(\mathcal{D},\mathcal{F})$ is good, specifically, that*
(i) *The ϵ-risk of an affine estimate can be made arbitrarily close to the quantity*

$$\Phi_+(\epsilon) = \inf_{\phi\in\mathcal{F},\alpha>0}\sup_{x,y\in X}\Phi_{\ln\left(\frac{2}{\epsilon}\right)}(x,y;\phi,\alpha)$$

(cf. Lemma 3.1), and
(ii) *We have* $\mathrm{Risk}_*(\epsilon) \ge \Phi_-(\epsilon) = \sup_{x,y\in X}\inf_{\phi\in\mathcal{F},\alpha>0}\Phi_{\frac{1}{2}\ln\left(\frac{1}{4\epsilon}\right)}(x,y;\phi,\alpha)$ *(cf. Lemma 3.2).*

As it is seen from the proofs of Lemmas 3.1, 3.2, both these statements hold true without the goodness assumption. The role of the latter is in ensuring that $\Phi_+(\epsilon)$ is within an absolute constant factor of $\Phi_-(\epsilon)$.

Lemma 3.2 Implies the following result:

PROPOSITION 3.1. *Under the premise of Theorem 3.1, the Hellinger affinity*

$$\mathrm{AffH}(\mu,\nu) = \int_\Omega \sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)$$

is a continuous and log-concave function on $\mathcal{M}\times\mathcal{M}$, and the quantity $\Phi_*(r)$, $r \ge 0$, admits the following representation:
(3.10)

$$2\Phi_*(r) = \max_{x,y}\left\{g^Tx - g^Ty : \mathrm{AffH}(A(x),A(y)) \ge \exp\{-r\}, x, y \in X\right\}.$$

We see that the upper bound $\Phi_*(\ln(2/\epsilon))$ on RiskAff(ϵ) stated in Theorem 3.1 admits a very transparent interpretation: this bound is the maximum of the variation $\frac{1}{2}\max_{x,y}[g^Tx - g^Ty]$ of the estimated functional on the set of pairs $x, y \in X$ with the associated distributions "close" to each other, namely, such that $\mathrm{AffH}(A(x), A(y)) \geq \epsilon/2$.

PROOF. By exactly the same argument as in the proof of Theorem 3.1, the function

$$\Psi(\mu,\nu;\phi) \quad = \quad [\ln\left(\int \exp\{-\phi(\omega)\}p_\mu(\omega)P(d\omega)\right) + \ln\left(\int \exp\{\phi(\omega)\}p_\nu(\omega)P(d\omega)\right)] : \\ (\mathcal{M}\times\mathcal{M})\times\mathcal{F} \to \mathbb{R}$$

is a well-defined and continuous on its domain, and this function is convex in ϕ and concave in (μ,ν). We claim that

$$\ln(\mathrm{AffH}(\mu,\nu)) = \frac{1}{2}\min_\phi \Psi(\mu,\nu;\psi), \tag{3.11}$$

which would imply that $\ln(\mathrm{AffH}(\cdot))$ is indeed a finite concave function on $\mathcal{M}\times\mathcal{M}$ and as such is continuous (recall that $\mathcal{M}$ is open). To justify our claim, note that for fixed $\mu,\nu \in \mathcal{M}$, setting $\phi = \frac{1}{2}\ln(p_\nu/p_\mu)$, we get a function from $\mathcal{F}$ such that $\Psi(\mu,\nu;\bar\phi) = 2\ln(\mathrm{AffH}(\mu,\nu))$. To complete the verification of (3.11), it suffices to demonstrate that $\Psi(\mu,\nu;\phi) \geq \Psi(\mu,\nu;\bar\phi)$ whenever $\phi\in\mathcal{F}$, which is immediate, since setting $\phi = \bar\phi + \Delta$, we have

$$\begin{aligned}
&\exp\{\Psi(\mu,\nu;\bar\phi)/2\} = \int\sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\\
&= \int\left[(p_\mu(\omega)p_\nu(\omega))^{1/4}\exp\{-\Delta(\omega)/2\}\right]\left[(p_\mu(\omega)p_\nu(\omega))^{1/4}\exp\{\Delta(\omega)/2\}\right]P(d\omega)\\
&\leq \left[\int\sqrt{p_\mu(\omega)p_\nu(\omega)}\exp\{-\Delta(\omega)\}P(d\omega)\right]^{1/2}\left[\int\sqrt{p_\mu(\omega)p_\nu(\omega)}\exp\{\Delta(\omega)\}P(d\omega)\right]^{1/2}\\
&= \exp\{\Psi(\mu,\nu;\phi)/2\}.
\end{aligned}$$

Now note that by (3.3)

$$\begin{aligned}
&2\Phi_*(r) = \sup_{x,y\in X}\left\{\inf_{\phi\in\mathcal{F},\alpha>0}\left[g^Tx - g^Ty + \alpha\Psi(A(x),A(y);\alpha^{-1}\phi) + 2\alpha r\right]\right\}\\
&= \sup_{x,y\in X}\left\{g^Tx - g^Ty + \inf_{\alpha>0}\alpha\left[\inf_{\phi\in\mathcal{F}}\Psi(A(x),A(y);\alpha^{-1}\phi) + 2r\right]\right\}\\
&= \sup_{x,y\in X}\left\{g^Tx - g^Ty + \inf_{\alpha>0}\alpha\left[\inf_{\psi\equiv\alpha^{-1}\phi\in\mathcal{F}}\Psi(A(x),A(y);\psi) + 2r\right]\right\}\\
&= \sup_{x,y\in X}\{g^Tx - g^Ty + \underbrace{\inf_{\alpha>0}\alpha\left[2\ln(\mathrm{AffH}(A(x),A(y))) + 2r\right]}_{=\begin{cases}0, & \ln(\mathrm{AffH}(A(x),A(y))) + r \geq 0\\ -\infty, & \ln(\mathrm{AffH}(A(x),A(y))) + r < 0\end{cases}}\} \quad [\text{see (3.11)}]\\
&= \max_{x,y}\left\{g^Tx - g^Ty : \mathrm{AffH}(A(x),A(y)) \geq \exp\{-r\}, x, y\in X\right\}
\end{aligned}$$

∎

3.2. *The case of multiple observations.* In Problem I, our goal was to estimate g^Tx from a *single* observation ω of the associated with x random variable $\omega \sim p_{A(x)}(\cdot)$. The result can be immediately extended to the case when we want to recover g^Tx from a sample of independent observations $\omega_1, ..., \omega_L$ of random variables ω_ℓ with distributions parameterized by x. Specifically, let (Ω_ℓ, P_ℓ) and $(\mathcal{D}_\ell, \mathcal{F}_\ell)$, $1 \le \ell \le L$, be as in Example 4, and let every pair $(\mathcal{D}_\ell, \mathcal{F}_\ell)$ be good. Assume, further, that $X \subset \mathbb{R}^n$ is a convex compact set and $A_\ell(x)$ are affine mappings with $A_\ell(X) \subset \mathcal{M}_\ell$. Given a linear form g^Tz on $\mathbb{R}^n$ and a sequence of independent realizations $\omega_\ell \sim p^\ell_{A_\ell(x)}(\cdot)$, $\ell = 1, ..., L$, we want to recover from these observations the value g^Tx of the given affine form at the "signal" x underlying our observations.

In our current situation, we call a candidate estimate $\widehat{g}(\omega_1, ..., \omega_L)$ *affine*, if it is of the form

$$\widehat{g}(\omega_1, ..., \omega_L) = \sum_{\ell=1}^{L} \phi_\ell(\omega_\ell), \tag{3.12}$$

where $\phi_\ell \in \mathcal{F}_\ell$, $\ell = 1, ..., M$. Note that setting $(\mathcal{D}, \mathcal{F}) = \bigotimes_{\ell=1}^{L} (\mathcal{D}_\ell, \mathcal{F}_\ell)$, we reduce the situation to the one we have already considered. In particular, Theorem 3.1 along with the proof of Lemma 3.1 implies the following result (where the ϵ-risks – of an estimate, the minimax optimal and the affine-minimax optimal – are defined exactly as in the single-observation case):

THEOREM 3.2. *In the just described situation, for $r > 0$ let*

(3.13)

$$\begin{array}{rcl} \Phi_r(x, y; \phi, \alpha) & = & \alpha\Big[\sum_{\ell=1}^{L} \ln\Big(\int_{\Omega_\ell} \exp\{-\alpha^{-1}\phi_\ell(\omega_\ell)\} p^\ell_{A_\ell(x)}(\omega_\ell P(d\omega_\ell)\Big) \\ & & + \Big(\int_{\Omega_\ell} \exp\{\alpha^{-1}\phi_\ell(\omega_\ell)\} p^\ell_{A_\ell(y)}(\omega_\ell P(d\omega_\ell)\Big)\Big] \\ & & + g^Tx - g^Ty + 2\alpha r : Z \times \mathcal{F}_+ \to \mathbb{R}, \\ Z & = & X \times X, \\ \mathcal{F}_+ & = & \mathcal{F}_1 \times ... \times \mathcal{F}_L \times \{\alpha > 0\}. \end{array}$$

The function Φ_r is continuous on its domain, concave in the (x, y)-argument and convex in the (ϕ, α)-argument and possesses a well defined saddle point value

(3.14)

$$2\Phi_*(r) = \sup_{x,y\in X} \underbrace{\inf_{\phi,\alpha\in\mathcal{F}_+} \Phi_r(x, y; \phi, \alpha)}_{\underline{\Phi}_r(x,y)} = \inf_{(\phi,\alpha)\in\mathcal{F}_+} \underbrace{\sup_{x,y\in X} \Phi_r(x, y; \phi, \alpha)}_{\overline{\Phi}_r(\phi,\alpha)}$$

which is a concave a nonnegative function of $r \geq 0$. Moreover,

(i) *For all $\epsilon \in (0, 1/4)$ we have*

$$\mathrm{RiskA}(\epsilon) \leq \Phi_*(\ln(2/\epsilon)) \leq \theta(\epsilon)\,\mathrm{Risk}_*(\epsilon),\ \theta(\epsilon) = \frac{2\ln(\frac{2}{\epsilon})}{\ln(\frac{1}{4\epsilon})}; \tag{3.15}$$

(ii) *Given $\epsilon \in (0, 1/4)$ and $\delta \geq 0$, in order to build an affine estimate with ϵ-risk not exceeding $[\Phi_*(\ln(2/\epsilon)) + \delta]$, where $\delta > 0$ is given, it suffices to find $\alpha_* > 0$ and $\phi_\ell^* \in \mathcal{F}_\ell$, $1 \leq \ell \leq L$ such that*

$$\overline{\Phi}_{\ln(2/\epsilon)}(\phi^*, \alpha_*) \leq 2\Phi_*(\ln(2/\epsilon)) + \delta/2 \tag{3.16}$$

to compute the quantity

$$\begin{array}{rcl} c & = & \frac{1}{2}\max_{x\in X}\Big[g^T x \\ & & + \alpha_* \sum_{\ell=1}^L \ln\Big(\int_{\Omega_\ell} \exp\{-\alpha^{-1}\phi_\ell^*(\omega_\ell)\} p^\ell_{A_\ell(x)}(\omega_\ell) P_\ell(d\omega_\ell)\Big)\Big] \\ & & -\frac{1}{2}\max_{y\in X}\Big[-g^T y \\ & & + \alpha_* \sum_{\ell=1}^L \ln\Big(\int_{\Omega_\ell} \exp\{\alpha^{-1}\phi_\ell^*(\omega_\ell)\} p^\ell_{A_\ell(y)}(\omega_\ell) P_\ell(d\omega_\ell)\Big)\Big] \end{array} \tag{3.17}$$

and to set

$$\widehat{g}(\omega_1, ..., \omega_L) = \sum_{\ell=1}^L \phi_\ell^*(\omega_\ell) + c. \tag{3.18}$$

REMARK 3.2. *Computing the "nearly optimal" affine estimate (3.18) reduces to Convex Programming and thus can be carried out efficiently, provided that we are given explicit descriptions of*

- *the linear spaces $\mathcal{F}_\ell$, $\ell = 1, ..., L$ (as it is the case, e.g., in Examples 1 – 3),*
- *X (e.g., by a list of efficiently computable convex constraints which cut X off $\mathbb{R}^n$)*

and are capable to compute efficiently the value of Φ_r at a point.

REMARK 3.3. *Assume that the observations ω_ℓ, $\ell_0 \leq \ell \leq \ell_1$, are copies of the same random variable (that is, $\Omega_\ell, P_\ell, \mathcal{D}_\ell, \mathcal{F}_\ell, A_\ell(\cdot)$ are independent of ℓ for $\ell_0 \leq \ell \leq \ell_1$). Then the convex function $\overline{\Phi}_r(\phi_1, ..., \phi_L, \alpha)$ is symmetric w.r.t. the arguments $\phi_\ell \in \mathcal{F}_{\ell_0}$, $\ell_0 \leq \ell \leq \ell_1$, and therefore when building the estimate (3.18) we lose nothing when restricting ourselves with ϕ's satisfying $\phi_\ell = \phi_{\ell_0}$, $\ell_0 \leq \ell \leq \ell_1$, which allows to reduce the computational effort of building α_*, ϕ_ℓ^*.*

3.2.1. *Illustration.* Consider the toy problem where we want to recover the probability p of getting 1 from a Bernoulli distribution, given L independent realizations $\omega_1, ..., \omega_L$ of the associated random variable. To handle the problem, we specialize our general setup as follows:

- (Ω_ℓ, P_ℓ), $1 \le \ell \le L$, are identical to the two-point set $\{0; 1\}$ with the counting measure;
- $\mathcal{M}$ is the interval $(0, 1)$, and $p_\mu(1) = 1 - p_\mu(0) = \mu$, $\mu \in \mathcal{M}$;
- X is a compact convex subset in $\mathcal{M}$, say, the segment $[1.e-16, 1-1/e-16]$, and $A(x) = x$.

Invoking Remark 3.3, we lose nothing when restricting ourselves with affine estimates of the form (3.12) with identical to each other functions $\phi_\ell(\cdot)$, $1 \le \ell \le L$, that is, with the estimates

$$\widehat{g}(\omega_1, ..., \omega_L) = \gamma + \delta \sum_{\ell=1}^{L} \omega_\ell.$$

Invoking Theorem 3.2, the coefficients γ and δ are readily given by the ϕ-component of the saddle point (max in $x, y \in X$, min in $\phi = [\phi_0; \phi_1] \in \mathbb{R}^2$ and $\alpha > 0$) of the convex-concave function

$$x - y + \alpha \left[L \ln\left(\epsilon^{-\phi_0/\alpha}(1-x) + \epsilon^{-\phi_1/\alpha} x\right) + L \ln\left(\epsilon^{\phi_0/\alpha}(1-y) + \epsilon^{\phi_1/\alpha} y\right) + 2\ln(2/\epsilon) \right];$$

the (guaranteed upper bound on the) ϵ-risk of this estimate is half of the corresponding saddle point value. The saddle point (it is easily seen that it does exist) can be computed to a whatever high accuracy by the standard Convex Programming techniques. In Table 1, we present the nearly optimal affine estimates along with the corresponding risks. In the table, the upper risk bound is the one guaranteed by Theorem 3.2, the "lower risk bound" is the largest d such that the hypotheses "$p = 0.5 + d$" and "$p = 0.5 - d$" cannot be distinguished from L independent observations of a random variable $\sim$ Bernoulli(p) with the sum of probabilities of errors $< 2\epsilon$ (this easily computable quantity is a lower bound on the minimax optimal ϵ-risk $\mathrm{Risk}_*(\epsilon)$), and $\vartheta(\epsilon) = \frac{2\ln(2/\epsilon)}{\ln(0.25/\epsilon)}$ is the theoretical upper bound on the "level of nonoptimality" of our estimate. As it could be guessed in advance, for large L the near-optimal affine estimate is close to the trivial estimate $\frac{1}{L}\sum_{\ell=1}^{L} \omega_\ell$.

4. Applications.

In this Section, we present some applications of Theorems 3.1, 3.2.

ϵ	L	γ	δ	upper risk bound	lower risk bound	ratio of bounds	$\vartheta(\epsilon)$
0.05	10	2.91e-1	4.18e-2	3.61e-1	2.49e-1	1.45	4.58
0.05	100	4.13e-2	9.17e-3	1.33e-1	8.19e-2	1.63	4.58
0.05	1000	4.29e-3	9.91e-4	4.29e-3	2.60e-3	1.65	4.58
0.01	10	3.58e-1	2.83e-2	4.04e-1	3.29e-1	1.23	3.29
0.01	100	5.83e-2	8.84e-2	1.59e-1	1.15e-1	1.38	3.29
0.01	1000	6.15e-3	9.88e-4	5.13e-2	3.67e-3	1.40	3.29
0.001	10	4.19e-1	1.61e-2	4.42e-1	3.98e-1	1.11	2.75
0.001	100	8.15e-2	8.37e-3	1.88e-1	1.51e-1	1.24	2.75
0.001	1000	8.79e-3	9.82e-4	6.14e-3	4.88e-3	1.26	2.75

TABLE 1
Recovering the parameter of a Bernoulli distribution.

4.1. *Positron Emission Tomography.* The Positron Emission Tomography (PET) is a non-invasive diagnostic tool allowing to visualize not only the anatomy of tissues in a body, but their functioning as well. In Pet, a patient is administered a radioactive tracer chosen in such a way that it concentrates in the areas of interest (e.g., those of high metabolic activity in early diagnosis of cancer). The tracer disintegrates emitting positrons which then annihilate with near-by electrons to produce pairs of photons flying at the speed of light in opposite directions; the orientation of the resulting *line of response* (LOR) is completely random. The patient is placed in a cylinder with the surface split into small cells – detectors. When two of detectors are hit by photons "nearly simultaneously" – within an appropriately chosen short time window, this event indicates that somewhere at a line crossing the detectors a disintegration act took place. Such an event is registered, and the data collected by the PET device form a list of the number of events registered in every one of the *bins* (pairs of detectors) in course of a given time t. The goal of a PET reconstruction algorithm is ro recover from this data the density of the tracer. The standard mathematical model of PET is as follows. After discretization of the field of view, there are N voxels (small 3D cubes) assigned with nonnegative (and unknown) amounts x_i of the traces, $i = 1, ..., n$. The number of LORs emanating from a voxel i is a realization of a Poisson random variable with parameter x_i, and these variables for different voxels are independent. Every LOR emanating from a voxel i is subject to a "lottery" which decides in which bin (pair of detectors) it will be registered, if it will be registered at all – some LOR's can intersect the surface of the cylinder only in one point or not intersect it at all and thus are missed. The role of lottery is played by random orientation of the LOR in question,

and outcomes of different lotteries are independent. The probabilities $q_{i\ell}$ for a LOR emanating from voxel i to be registered in bin ℓ are known (they are readily given by the geometry of the device). With this model, the data registered by PET is a realization of a random vector $(\omega_1, ..., \omega_L)$ (M is the total number of bins) with independent Poisson-distributed coordinates, the parameter of the Poisson distribution associated with ω_ℓ being

$$A_\ell(x) = \sum_{i=1}^{n} q_{i\ell} x_i.$$

Assume that our a priori information on x allows to point out a convex compact set $X \subset \{x \in \mathbb{R}^n : x > 0\}$ such that $x \in X$. Assuming w.l.o.g. that $\sum_i q_{i\ell} > 0$ for every ℓ (indeed, we can eliminate all bins ℓ which never register LORs) and invoking Example 2, we find ourselves in the situation of Section 3.2. It follows that in order to evaluate a given linear form $g^T x$ of the unknown tracer density x, we can use the construction from Theorem 3.2 to build a near-optimal affine estimate of $g^T x$. The recipe suggested to this end by Theorem 3.2 reads as follows: the estimate is of the form

$$\widehat{g}(\omega) = \sum_{\ell=1}^{L} \gamma_\ell^* y_\ell + c_*, \tag{4.1}$$

where y_ℓ is the number of LORs registered in bin ℓ and $\gamma^* = [\gamma_1^*; ...; \gamma_L^*]$, c_* are given by an optimal solution (γ^*, α_*) to the convex optimization problem

(4.2)

$$\begin{array}{rcl} & & \min\limits_{\alpha>0,\gamma} \overline{\Phi}_r(\gamma, \alpha), \\ \overline{\Phi}_r(\gamma, \alpha) & = & \max\limits_{x,y \in X} \Big\{ g^T x - g^T y \\ & & \quad + \alpha \Big[\sum_{\ell=1}^{L} \left[q_\ell(x) \exp\{-\alpha^{-1}\gamma_\ell\} + q_\ell(y) \exp\{\alpha^{-1}\gamma_\ell\} \right] \\ & & \qquad - q(x) - q(y) + 2r \Big] \Big\}, \\ & & r = \ln(2/\epsilon),\ q_\ell(z) = \sum_{i=1}^{n} q_{i\ell} z_i,\ q(z) = \sum_{\ell=1}^{L} q_\ell(z). \end{array}$$

(it is easily seen that the problem is solvable), with

(4.3)

$$\begin{array}{rcl} c_* & = & \frac{1}{2} \Big[\max_{x \in X} \left\{ g^T x + \alpha_* \left[-q(x) + \sum_{\ell=1}^{L} q_\ell(x) \exp\{-\alpha_*^{-1}\gamma_\ell^*\} \right] \right\} \\ & & \quad - \max_{y \in X} \left\{ -g^T y + \alpha_* \left[-q(y) + \sum_{\ell=1}^{L} q_\ell(y) \exp\{\alpha_*^{-1}\gamma_\ell^*\} \right] \right\} \Big]. \end{array}$$

4.2. *Gaussian observations.* Now consider the standard problem of recovering a linear form $g^T x$ of a signal x known to belong to a given convex compact set $X \subset \mathbb{R}^n$ via indirect observations of the signal corrupted by Gaussian noise. W.l.o.g., let the model of observations be

$$\omega = Ax + \xi,\ \xi \sim \mathcal{N}(0, I_L). \tag{4.4}$$

The associated pair $(\mathcal{D}, \mathcal{F})$ is comprised of the shifts of the standard Gaussian distribution $(\mathcal{D})$ and all affine forms on $\mathbb{R}^L$ $(\mathcal{F})$ and is good (see Example 3). The affine estimates in the case in question are just the affine functions of ω. The near-optimality of affine estimates in the case in question was established by D. Donoho [5], not only for the ϵ-risk, but for all risks based on the standard loss functions. We are about to augment the results of Donoho by the following result on asymptotic optimality of affine estimates:

PROPOSITION 4.1. *In the situation in question, the affine estimate $\widehat{g}_\epsilon(\cdot)$ yielded by Theorem 3.2 is asymptotically ($\epsilon \to +0$) optimal, specifically,*

$$\begin{aligned} \epsilon \in (0, 1/2) \Rightarrow \mathrm{Risk}(\widehat{g}_\epsilon; \epsilon) &\leq \psi(\epsilon)\, \mathrm{Risk}_*(\epsilon), \\ \psi(\epsilon) &= \frac{\sqrt{2\ln(2/\epsilon)}}{\mathrm{ErfInv}(\epsilon)} = 1 + o(1) \ \text{as } \epsilon \to +0. \end{aligned} \tag{4.5}$$

PROOF. Let $G(\cdot)$ be the density of the $\mathcal{N}(0, I_L)$ distribution. By Theorem 3.2, we have $\mathrm{Risk}(\widehat{g}_\epsilon; \epsilon) \leq \Phi_*(\ln(2/\epsilon))$, where for $r > 0$

$$\begin{aligned} 2\Phi_*(r) &= \max_{x,y \in X} \underline{\Phi}_r(x, y), \\ \underline{\Phi}_r(x, y) &= \inf_{\phi \in \mathbb{R}^L, \alpha > 0} \Big\{ g^T x - g^T y \\ &\quad + \alpha \Big[\ln\Big(\int \exp\{-\alpha^{-1} \phi^T \omega\} G(\omega - Ax) d\omega \Big) \\ &\qquad + \ln\Big(\int \exp\{\alpha^{-1} \phi^T \omega\} G(\omega - Ay) d\omega \Big) + 2r \Big] \Big\} \\ &= \inf_{\phi \in \mathbb{R}^L, \alpha > 0} \Big\{ g^T x - g^T y + \phi^T A(y - x) + 2\Big[\alpha^{-1} \tfrac{\phi^T \phi}{2} + \alpha r \Big] \Big\} \\ &= \inf_{\phi} \Big\{ g^T x - g^T y + \phi^T A(x - y) + 2\sqrt{2r} \|\phi\|_2 \Big\} \\ &= \begin{cases} g^T x - g^T y, & \|A(x - y)\|_2 \leq 2\sqrt{2r} \\ -\infty, & \|A(x - y)\|_2 > 2\sqrt{2r} \end{cases} \end{aligned}$$

Thus,

$$\mathrm{Risk}(\widehat{g}_\epsilon; \epsilon) \leq \Phi_*(\ln(2/\epsilon)) = \frac{1}{2}[g^T \bar{x} - g^T \bar{y}] \tag{4.6}$$

for certain $\bar{x}, \bar{y} \in X$ with $\|A(x-y)\|_2 \leq 2\sqrt{2\ln(2/\epsilon)}$. It remains to prove that

$$\mathrm{Risk}_*(\epsilon) \geq \psi^{-1}(\epsilon)\frac{1}{2}\Phi_*(\ln(2/\epsilon)). \tag{4.7}$$

To this end assume, on the contrary to what should be proved, that

$$\mathrm{Risk}_*(\epsilon) < \psi^{-1}(\epsilon)\Phi_*(\ln(2/\epsilon)) \qquad [= \psi^{-1}(\epsilon)\tfrac{1}{2}[g^T\bar{x} - g^T\bar{y}]]$$

and let us lead this assumption to a contradiction. Under our assumption, there exists $\rho < \psi^{-1}(\epsilon)\frac{1}{2}[g^T\bar{x} - g^T\bar{y}]$, $\epsilon' < \epsilon$ and an estimate $\widetilde{g}$ such that

$$\forall(x \in X) : \mathrm{Prob}\{|\widetilde{g}(Ax+\xi) - g^T x| \geq \rho\} \leq \epsilon'. \tag{4.8}$$

Observing that $\psi(\epsilon) > 1$, we see that $2\rho < [g^T\bar{x} - g^T\bar{y}]$. Let $\widehat{x} = \bar{x}$ and $\widehat{y}$ be a convex combination of $\bar{x}$ and $\bar{y}$ such that $2\rho = [g^T\widehat{x} - g^T\widehat{y}]$. Note that

$$\|A(\widehat{x}-\widehat{y})\|_2 = \underbrace{\left[\frac{2\rho}{[g^T\bar{x} - g^T\bar{y}]}\right]}_{<\psi^{-1}(\epsilon)} \|A(\bar{x}-\bar{y}\|_2 \leq \psi^{-1}(\epsilon)2\sqrt{2\ln(2/\epsilon)} = 2\,\mathrm{erfinv}(\epsilon).$$

Now let Π_1 be the hypothesis that the distribution of an observation (4.4) comes from $x = \widehat{x}$, and let Π_2 be the hypothesis that this distribution comes from $x = \widehat{y}$. From (4.8) by the same standard argument as in the proof of Lemma 3.2 it follows that there exists a routine, based on a single observation (4.4), for distinguishing between Π_1 and Π_2 which rejects Π_i when this hypothesis is true with probability $\leq \epsilon'$, $i = 1, 2$. But is is well known when the hypotheses on shifts of the standard Gaussian distribution indeed can be distinguished with the outlined reliability: this is possible if and only if the Euclidean distance between the corresponding shifts is at least $2\,\mathrm{erfinv}(\epsilon')$. This condition is *not* satisfied for our Π_i, $i = 1, 2$ which correspond to shifts $A\widehat{x}$, $A\widehat{y}$, since $\|A\widehat{x} - A\widehat{y}\|_2 \leq 2\,\mathrm{erfinv}(\epsilon) < 2\,\mathrm{erfinv}(\epsilon)$. We have arrived at a desired contradiction. ∎

In fact, the reasoning can be slightly simplified and strengthened to yield the following result:

PROPOSITION 4.2. *In the situation of Proposition 4.1, one can build efficiently an affine estimate $\widehat{g}_\epsilon$ such that*

$$0 < \epsilon < 1/2 \Rightarrow \mathrm{Risk}(\widehat{g}_\epsilon; \epsilon) \leq \frac{\mathrm{ErfInv}(\epsilon/2)}{\mathrm{ErfInv}(\epsilon)}\,\mathrm{Risk}_*(\epsilon) \tag{4.9}$$

(cf. Proposition 4.1 and note that $\frac{\mathrm{ErfInv}(\epsilon/2)}{\mathrm{ErfInv}(\epsilon)} < \frac{\sqrt{2\ln(2/\epsilon)}}{\mathrm{ErfInv}(\epsilon)}$*).*

PROOF. Let

$$\Psi(x,y;\phi) = g^T x - g^T y + \phi^T A(y-x) + 2\,\mathrm{erfinv}(\epsilon/2)\|\phi\|_2 : (X\times X)\times\mathbb{R}^L \to \mathbb{R}.$$

Ψ clearly is a continuous convex in ϕ and concave in (x,y) function on its domain; by the same argument as in the proof of Theorem 3.1, Ψ has a well-defined saddle point value

$$2\Psi_*(\epsilon) = \inf_\phi \overbrace{\max_{x,y\in X}\Psi(x,y;\phi)}^{\overline{\Psi}(\phi)} = \max_{x,y\in X}\overbrace{\inf_\phi \Psi(x,y;\phi)}^{\underline{\Psi}(x,y)}.$$

The function

$$\overline{\Psi}(\phi) = \max_{x,y\in X}\Big[g^T x - g^T y + \phi^T(Ay - Ax)\Big] + 2\,\mathrm{erfinv}(\epsilon/2)\|\phi\|_2 \geq 2\,\mathrm{erfinv}(\epsilon)\|\phi\|_2$$

is a finite convex function on $\mathbb{R}^L$ which goes to ∞ as $\|\phi\|_2 \to \infty$ and therefore it attains its minimum at a point ϕ_*, so that

$$2\Psi_*(\epsilon) = \overline{\Psi}(\phi_*).$$

Setting

$$c_* = \frac{1}{2}\left[\max_{x\in X}[g^T x - \phi_*^T Ax] - \max_{y\in Y}[-g^T y + \phi_*^T Ay]\right],$$

we, same as in the proof of Lemma 3.1, have

$$\begin{array}{ll}(a) & \max_{x\in X}\left[g^T x - \phi_*^T Ax - c_*\right] + \mathrm{erfinv}(\epsilon/2)\|\phi_*\|_2 = \Psi_*(\epsilon),\\ (b) & \max_{y\in X}\left[-g^T y + \phi_*^T Ax + c_*\right] + \mathrm{erfinv}(\epsilon/2)\|\phi_*\|_2 = \Psi_*(\epsilon).\end{array}$$

Now consider the affine estimate

$$\widehat{g}_\epsilon(\omega) = \phi_*^T\omega + c_*.$$

From (a) it follows that

$$\forall d > \Psi_*(\epsilon) : \sup_{x\in X}\mathrm{Prob}\left\{g^T x - \widehat{g}_\epsilon(Ax+\xi) > d\right\} \leq \epsilon' < \epsilon/2,$$

while (b) implies that

$$\forall d > \Psi_*(\epsilon) : \sup_{y\in X}\mathrm{Prob}\left\{\widehat{g}_\epsilon(Ay+\xi) - g^T y > d\right\} \leq \epsilon' < \epsilon/2.$$

We conclude that $\mathrm{Risk}(\widehat{g}_\epsilon;\epsilon) \le \Psi_*(\epsilon)$. To complete the proof, it suffices to demonstrate that

$$\mathrm{Risk}_*(\epsilon) \le \frac{\mathrm{ErfInv}(\epsilon/2)}{\mathrm{ErfInv}(\epsilon)}\Psi_*(\epsilon). \tag{4.10}$$

To this end, observe that

$$\begin{array}{rcl}\underline{\Psi}(x,y) & = & \left[g^Tx - g^Ty\right] + \inf_\phi\left\{\phi^TA(y-x) + 2\,\mathrm{erfinv}(\epsilon/2)\|\phi\|_2\right\}\\ & = & \left\{\begin{array}{ll} g^Tx - g^Ty, & \|A(y-x)\|_2 \le 2\,\mathrm{erfinv}(\epsilon/2)\\ -\infty, & \text{otherwise}\end{array}\right.,\end{array}$$

whence

$$\mathrm{Risk}_*(\widehat{g}_\epsilon;\epsilon) \le \Psi_*(\epsilon) = \frac{1}{2}[g^T\bar{x} - g^T\bar{y}] \tag{4.11}$$

for certain $\bar{x},\bar{y} \in X$ such that $\|A(\bar{x}-\bar{y})\|_2 \le 2\,\mathrm{erfinv}(\epsilon)$. Relation (4.10) can be derived from this observation by exactly the same argument as used in the proof of Proposition 4.1 to derive (4.7) from (4.6). ∎

5. Adaptive version of the estimate. In the situation of Problem I, let $X^1 \subset X^2 \subset ... \subset X^K$ be a nested collection of nonempty convex compact sets in $\mathbb{R}^n$ such that $A(X^K) \subset \mathcal{M}$. Consider a modification of the problem where the signal x underlying our observation is known to belong to one of X^k with unknown in advance value of $k \le K$. Given a linear form g^Tz on $\mathbb{R}^n$, let $\mathrm{Risk}^k(\widehat{g};\epsilon)$ and $\mathrm{Risk}^k_*(\epsilon)$ be, respectively, the ϵ-risk of an estimate $\widehat{g}$ on X^k, and the minimax optimal ϵ-risk of recovering g^Tx on X^k. Let also $\Phi^k_*(r)$ be the function associated with $X = X^k$ according to (3.3). As it is immediately seen, the functions $\Phi^k_*(r)$ grow with k. Our goal is to modify the estimate $\widehat{g}$ yielded by Theorem 3.1 in such a way that the ϵ-risk of the modified estimate on X^k will be "nearly" $\mathrm{Risk}^k_*(\epsilon)$ *for every* $k \le K$. This goal can be achieved by a straightforward application of the well-known Lepskii's adaptation scheme [20, 21] as follows.

Given $\delta > 0$, let $\delta' \in (0,\delta)$, and let $\widehat{g}^k(\cdot)$ be the affine estimate with the (ϵ/K)-risk on X^k not exceeding $\Phi^k_*(\ln(2K/\epsilon)) + \delta'$ provided by Theorem 3.1 as applied with ϵ/K substituted for ϵ and X^k substituted for X. Then, for any $k \le K$

$$\begin{array}{rl}\sup\limits_{x\in X^k} & \mathrm{Prob}_{\omega\sim p_{A(x)}(\cdot)}\left\{|\widehat{g}^k(\omega) - g^Tx| > \Phi^k_*(\ln(2K/\epsilon)) + \delta\right\}\\ & \le \epsilon'/K < \epsilon/K.\end{array} \tag{5.1}$$

Given observation ω, let us say that an index $k \le K$ is *ω-good*, if

(5.2)
$$\forall (k', k \le k' \le K) : |\widehat{g}^{k'}(\omega) - \widehat{g}^{k}(\omega)| \le \Phi_*^k(\ln(2K/\epsilon)) + \Phi_*^{k'}(\ln(2K/\epsilon)) + 2\delta.$$

Note that ω-good indexes do exist (e.g., $k = K$). Given ω, we can find the smallest ω-good index $k = k(\omega)$; our estimate is nothing but $\widehat{g}(\omega) = \widehat{g}^{k(\omega)}(\omega)$.

PROPOSITION 5.1. *Assume that $\epsilon \in (0, 1/4)$, and let*

$$\vartheta = 3\frac{\ln(2K/\epsilon)}{\ln(2/\epsilon)}. \tag{5.3}$$

Then

(5.4)
$$\forall (k, 1 \le k \le K) : \sup_{x \in X^k} \mathrm{Prob}_{\omega \sim p_{A(x)}(\cdot)} \left\{ |\widehat{g}(\omega) - g^T x| > \vartheta \Phi_*^k(\ln(2K/\epsilon)) + 3\delta \right\} < \epsilon,$$

whence also

$$\forall (k, 1 \le k \le K) : \mathrm{Risk}^k(\widehat{g}; \epsilon) \le \frac{6 \ln\left(\frac{2K}{\epsilon}\right)}{\ln\left(\frac{1}{4\epsilon}\right)} \mathrm{Risk}_*^k(\epsilon) + 3\delta. \tag{5.5}$$

PROOF. Setting $r = \ln(2K/\epsilon)$, let us fix $\bar{k} \le K$ and $x \in X^{\bar{k}}$ and call a realization ω *x-good*, if

$$\forall (k, \bar{k} \le k \le K) : |\widehat{g}^k(\omega) - g^T x| \le \Phi_*^k(r) + \delta. \tag{5.6}$$

Since $X^k \supset X^{\bar{k}}$ when $k \ge \bar{k}$, (5.1) implies that

$$\mathrm{Prob}_{\omega \sim p_{A(x)}(\cdot)} \{\omega \text{ is good } \} \ge 1 - \epsilon'. \tag{5.7}$$

Now, when x is the signal and ω is x-good, relations (5.6) imply that $\bar{k}$ is an ω-good index, so that $k(\omega) \le \bar{k}$. Since $k(\omega)$ is an ω-good index, we have

$$|\widehat{g}(\omega) - \widehat{g}^{\bar{k}}(\omega)| = |\widehat{g}^{k(\omega)}(\omega) - \widehat{g}^{\bar{k}}(\omega)| \le \Phi_*^k(r) + \Phi_*^{\bar{k}}(r) + 2\delta,$$

which combines with (5.6) to imply that

$$|\widehat{g}(\omega) - g^T x| \le 2\Phi_*^{\bar{k}}(r) + \Phi_*^{k(\omega)}(r) + 3\delta \le 3\Phi_*^{\bar{k}}(r) + 3\delta, \tag{5.8}$$

where the concluding inequality is due to $k(\omega) \le \bar{k}$ and to the fact that Φ_*^k grows with k. The bound (5.8) holds true whenever ω is x-good, which, as

we have seen, happens with probability $\geq 1-\epsilon'$. Since $\epsilon' < \epsilon$ and $\bar{x} \in X^{\bar{k}}$ is arbitrary, we conclude that

$$\mathrm{Risk}^{\bar{k}}(\widehat{g};\epsilon) \leq 3\Phi_*^{\bar{k}}(r) + 3\delta. \tag{5.9}$$

Using the nonnegativity and concavity of $\Phi_*^{\bar{k}}(\cdot)$ on the nonnegative ray and recalling the definition of r, we obtain $\Phi_*^k(r) \leq \frac{\ln(2K/\epsilon)}{\ln(2/\epsilon)}\Phi_*^k(\ln(2/\epsilon))$ whenever $\epsilon \leq 1/2$ and $K \geq 1$. Recalling the definition of ϑ, the right hand side in (5.9) does not therefore exceed $\vartheta\Phi_*^{\bar{k}}(\ln(2/\epsilon)) + 3\delta$. Since $\bar{k} \leq K$ is arbitrary, we have proved (5.4). This bound, due to Lemma 3.2, implies (5.5). ∎

6. From estimating linear forms to signal recovery. We have seen that in the context of Problem I we know how to recover a linear form of the unknown signal with ϵ-risk just by an absolute constant factor larger than the minimax optimal risk. Our goal is to demonstrate that when X has a favorable geometry, nearly optimal estimates of linear forms imply "not too bad" estimates of the unknown signal. For the sake of simplicity, we focus on recovery of the signal in the standard Euclidean norm. A candidate estimate is now a Borel function $\widehat{x}(w)$ taking values in the space $\mathbb{R}^n$ where x lives. Given a tolerance $\epsilon \in (0,1)$, we quantify the quality of such an estimate by the worst-case, over $x \in X$, upper $(1-\epsilon)$-quantile of the recovering error as measured in the Euclidean norm:

$$\mathrm{Risk}_2(\widehat{x};\epsilon) = \inf\left\{\delta : \sup_{x\in X} \mathrm{Prob}_{\omega\sim p_{A(x)}(\cdot)}\left\{\|\widehat{x}(\omega) - x\| > \delta\right\} < \epsilon\right\},$$

and denote by $\mathrm{Risk}_{2,*}(\epsilon)$ the associated minimax optimal risk:

$$\mathrm{Risk}_{2,*}(\epsilon) = \inf_{\widehat{x}(\cdot)} \mathrm{Risk}_2(\widehat{x};\epsilon).$$

Construction. Let us choose somehow a collection of N unit vectors $e_1, ..., e_N$ in $\mathbb{R}^n$. For $\epsilon \in (0, 0.1)$, let $\mathrm{Risk}_*^\ell(\epsilon)$ be the optimal, in the minimax sense, ϵ-risk of recovering $e_\ell^T x$ via an observation $\omega \sim p_{A(x)}(\cdot)$. Invoking Theorem 3.1, for a given $\epsilon \in (0, 0.1)$ is given, we can build estimates $\widehat{e}_\epsilon^\ell(\cdot)$ of the linear forms $e_\ell^T x$ and compute upper bounds $R^\ell(\epsilon)$ on the ϵ-risks of the estimates:

$$R^\ell(\epsilon) > \inf\left\{\delta : \sup_{x\in X} \mathrm{Prob}_{\omega\sim p_{A(x)}(\cdot)}\left\{|\widehat{e}_\epsilon^\ell(w) - e_\ell^T x| > \delta\right\} < \epsilon\right\}$$

in such a way that R^ℓ are within an absolute constant factor C of the minimax optimal ϵ-risks $\mathrm{Risk}_*^\ell(\epsilon)$ of recovering $e_\ell^T x$, $x \in X$:

$$R^\ell(\epsilon) < C\,\mathrm{Risk}_*^\ell(\epsilon). \tag{6.1}$$

Now, given $\bar{\epsilon} \in (0, 0.1)$, consider the following estimate $\widehat{x}$ of a signal $x \in X$ via observations ω. We build the N estimates $\widehat{e}^\ell(\cdot) \equiv \widehat{e}^\ell_{\bar{\epsilon}/N}(\cdot)$, $1 \le \ell \le N$. We further take as $\widehat{x}(\omega)$ (any) vector u satisfying the relations

$$u \in X \quad \text{and} \quad |e_\ell^T u - \widehat{e}^\ell_{\bar{\epsilon}/N}(\omega)| \le R^\ell(\bar{\epsilon}/N),\ \ell = 1, ..., N, \tag{6.2}$$

if such an u exists, otherwise $\widehat{x}(\xi^N)$ is a once for ever fixed point of X.

Analysis. Let $p_\infty(z) = \max_\ell |e_\ell^T z|$, $z \in \mathbb{R}^n$, and let $\mathrm{Risk}_{\infty,*}(\epsilon)$ be the optimal, in the minimax sense, ϵ-risk of recovering $x \in X$ via an observation $\omega \sim p_{A(x)}(\cdot)$, the loss function being $p_\infty(\cdot)$:

$$\begin{aligned}
\mathrm{Risk}_{\infty,*}(\epsilon) &= \inf_{\widetilde{x}(\cdot)} \mathrm{Risk}_\infty(\widetilde{x}; \epsilon), \\
\mathrm{Risk}_\infty(\widetilde{x}; \epsilon) &= \inf \left\{ \delta : \sup_{x \in X} \mathrm{Prob}_{\omega \sim p_{A(x)}(\cdot)} \left\{ p_\infty(\widetilde{x}(\xi^N) - x) > \delta \right\} < \epsilon. \right\}.
\end{aligned}$$

Since $\|\cdot\| \ge p_\infty(\cdot)$ due to $\|e_\ell\| = 1$, we have

$$\mathrm{Risk}_{\infty,*}(\epsilon) \le \mathrm{Risk}_{2,*}(\epsilon). \tag{6.3}$$

Our goal is to compare $\mathrm{Risk}_\infty(\widehat{x}; \bar{\epsilon})$ and $\mathrm{Risk}_{\infty,*}(\bar{\epsilon})$. By the origin of the estimates and R^ℓ, when $\omega \sim p_{A(x)}(\cdot)$ with $x \in X$, every one of the N inequalities $|e_\ell^T x - \widehat{e}^\ell_{\bar{\epsilon}/N}(\xi^N)| \le R^\ell(\bar{\epsilon}/N)$ takes place with probability $\ge 1 - \bar{\epsilon}/N + \delta$ with certain $\delta > 0$ independent of x. If all these inequalities take place (which happens with probability $\ge 1 - \bar{\epsilon} + \delta$), (6.2) is feasible (since the constraints in (6.2) are satisfied when $u = x$), and for every feasible solution u to (6.2) we have $|e_\ell^T u - e_\ell^T x| \le 2R^\ell(\bar{\epsilon}/N)$. Thus, we have

$$\sup_{x \in X} \mathrm{Prob}_{\omega \sim p_{A(x)}(\cdot)} \left\{ |e_\ell^T [\widehat{x}(\xi^N) - x]| \le 2R^\ell(\bar{\epsilon}/N),\ \ell = 1, ..., n \right\} < \bar{\epsilon},$$

whence

$$\mathrm{Risk}_\infty(\widehat{x}; \bar{\epsilon}) \le 2 \max_{1 \le \ell \le N} R^\ell(\bar{\epsilon}/N). \tag{6.4}$$

Now, due to the origin $R^\ell(\cdot)$ and to the fact that $\Phi_*(r)$ is nonnegative and concave function of $r \ge 0$, we have $R^\ell(\bar{\epsilon}/N) \le O(\ln(2N/\bar{\epsilon})/\ln(2/\bar{\epsilon}))R^\ell(\bar{\epsilon})$ (cf. the concluding step in the proof of Lemma 3.2). Thus, (6.4) implies that

$$\mathrm{Risk}_\infty(\widehat{x}; \bar{\epsilon}) \le \vartheta \max_{\ell \le N} R^\ell(\bar{\epsilon}),\ \vartheta = O(\ln(N/\bar{\epsilon})/\ln(1/\bar{\epsilon})),$$

which combines with (6.1) to imply that

$$\begin{gathered}
\mathrm{Risk}_\infty(\widehat{x}; \bar{\epsilon}) \le \widehat{\vartheta} \max_{\ell \le N} R^\ell_*(\bar{\epsilon}) \le \widehat{\vartheta}\, \mathrm{Risk}_{\infty,*}(\bar{\epsilon}) \le \widehat{\vartheta}\, \mathrm{Risk}_{2,*}(\bar{\epsilon}), \\
\widehat{\vartheta} = O(\ln(N/\bar{\epsilon})/\ln(1/\bar{\epsilon})).
\end{gathered} \tag{6.5}$$

Note that $\widehat{\vartheta}$ is a moderate constant, unless N is astronomically large. We conclude that *unless N is astronomically large, the estimate $\widehat{x}$ is nearly optimal in the sense of its ϵ-risk on X associated with the loss function p_∞.*

Now assume that the geometry of X allows to choose a collection $\{e_\ell\}$ of a "moderate" number N of unit vectors in such a way that

$$\forall u \in X - X : \|u\| \le C_X p_\infty^{\chi(X)}(u) \tag{6.6}$$

where $C_X > 0$ and $\chi(X) \in (0, 1]$ are appropriate constants. Since $\widehat{x}$ takes values in X, (6.6) combines with (6.5) to imply that

$$\mathrm{Risk}_2(\widehat{x}; \bar{\epsilon}) \le C_X \left[\mathrm{Risk}_\infty(\widehat{x}, \bar{\epsilon})\right]^{\chi(X)} \le C_X \widehat{\vartheta}^{\chi(X)} \left[\mathrm{Risk}_{2,*}(\bar{\epsilon})\right]^{\chi(X)}, \tag{6.7}$$

so that *the $\bar{\epsilon}$-risk of the estimate $\widehat{x}$, the loss function being $\|\cdot\|$, can be bounded from above in terms of the corresponding minimax optimal risk.* Ideally, we would like to have $\chi(X) = 1$, meaning that our estimate $\widehat{x}$ is "nearly minimax optimal" in terms of $\|\cdot\|$-risk (recall that for all practical purposes, $\widehat{\vartheta}$ is a moderate constant). How "far" we are from this ideal situation (that is, how far is $\chi(X)$ from 1), it depends solely on the geometry of X and is completely independent of how good is the affine mapping $A(x)$. It should be added that there are important situations where (6.6) is satisfied with "not too bad" constants C_X, $\chi(X)$. Here are two instructive examples:

Example 1: ℓ_1-ball. Assume that $X \subset \Delta_R = \{x \in \mathbb{R}^N : \sum_i |x_i| \le R\}$ (this is a frequently used model of a "sparse" signal). In this case, choosing as e_i the standard basic orths, we clearly have $\sum_i |u_i| \le 2R$ for every $u \in X - X$, whence $\|u\| = \sqrt{\sum_i u_i^2} \le \sqrt{p_\infty(u) \sum_i |u_i|} \le \sqrt{2R} p_\infty^{1/2}(u)$, that is, (6.6) holds true with $C_X = \sqrt{2R}$, $\chi(X) = 1/2$.

Example 2: Ellipsoid. Now assume that X is a centered at the origin ellipsoid with half-axes $d_i = Ri^{-\gamma}$, $\gamma > 0$ (this is the standard model of signals from Sobolev balls restricted onto uniform grids). In this case, assuming w.l.o.g. that the directions of the ellipsoid axes are the standard basic orths and choosing these orths as $e_1, ..., e_N$, for $u \in X - X$ we have

$$\sum_{i=1}^N u_i^2 i^{2\gamma} \le (2R)^2,$$

whence for every integer $k \ge 0$ one has $\sum_{i=k+1}^N u_i^2 \le (2R)^2 (k+1)^{-2\gamma}$. It follows that for every integer $k \ge 0$ we have

$$\|u\|^2 \le k p_\infty^2(u) + \sum_{i=k+1}^n u_i^2 \le k p_\infty^2(u) + (2R)^2 (k+1)^{-2\gamma}.$$

Minimizing the resulting bound in k, we get $\|u\| \leq O(1)R^{\frac{1}{2\gamma+1}}p_\infty^{\frac{2\gamma}{2\gamma+1}}(u)$, that is, in the case in question $C_X = O(1)R^{\frac{1}{2\gamma+1}}$, $\chi(X) = \frac{2\gamma}{2\gamma+1}$.

References.

[1] Ben-Tal, A., and Nemirovski, A., *Lectures on Modern Convex Optimization: Analysis, Algorithms and Engineering Applications*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2001.

[2] Birgé, L., Sur un théorème de minimax et son application aux tests. (French) *Probab. Math. Stat.* **3** (1982), 259-282.

[3] Birgé, L., Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **65** (1983), 181-237.

[4] Donoho, D., Liu, R., *Geometrizing rates of convergence, I*, Technical Report 137a, Dept. of Stat., University of California, Berkeley (1987).

[5] Donoho, D., Liu, R., Geometrizing Rates of Convergence, II. *The Annals of Statistics* **19:2** (1991), 633-667.

[6] Donoho, D., Liu, R., Geometrizing Rates of Convergence, III. *The Annals of Statistics* **19:2** (1991), 668-701.

[7] Donoho, D., Statistical estimation and optimal recovery. *The Annals of Statistics* **22:1** (1995), 238-270.

[8] Eubank, R., *Spline smoothing and Nonparametric Regression*, Dekker, New York, 1988.

[9] Goldenshluger, A., Nemirovski, A., On spatially adaptive estimation of nonparametric regression. *Math. Methods of Statistics*, **6:2** (1997), 135 – 170.

[10] Härdle, W., *Applied Nonparametric Regression.* ES Monograph Series 19, Cambridge, U.K., Cambridge University Press, 1990.

[11] Härdle, W., Kerkyacharian, G., Picard, D., Tsybakov, A.B., *Wavelets, Approximation and Statistical Applications.* Lecture Notes in Statistics **129**, Springer, New York, 1998.

[12] Hiriart-Urruty, J.B., Lemarechal, C. *Convex Analysis and Minimization Algorithms Vol. 1.* Springer Verlag, Berlin, 1993.

[13] Ibragimov, I.A., Khasminski, R.Z., *Statistical Estimation: Asymptotic Theory*, Springer, 1981.

[14] Ibragimov, I.A., Khas'minskij, R.Z., On the nonparametric estimation of a value of a linear functional in Gaussian white noise. (Russian. English summary) *Teor. Veroyatn. Primen.* **29**, No.1 (1984), 19-32.

[15] Ibragimov, I., Nemirovski, A., Khas'minski, R., Some problems of nonparametric estimation in Gaussian white noise. *Theory Probab. Appl.* **31:3** (1986), 391-406.

[16] Klemela, J., Tsybakov, A.B., Sharp adaptive estimation of linear functionals. *Annals of Statistics* **29** (2001), 1567-1600.

[17] Korostelev, A., Tsybakov, A., *Minimax theory of image reconstruction. Lecture Notes in Statistics* **82**, Springer, New York, 1993.

[18] Le Cam, L., Convergence of estimates under dimensionality restrictions, *Ann. Statist.* , **1** (1973), 38-53.

[19] Le Cam, L., *Asymptotic Methods in Statistical Decision Theory*, Springer, NY (1986).

[20] Lepskii, O., On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, **35:3** (1990), 454-466.

[21] Lepskii, O., Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory of Probability and Its Applications*, **36:4** (1991), 682-697.

[22] Lepskii, O., Mammen, E., Spokoiny, V., Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Annals of Statistics* **25:3** (1997), 929-947.
[23] Nemirovski, A., On nonparametric estimation of smooth regression functions. *Sov. J. Comput. Syst. Sci.*, **23:6** (1985), 1-11.
[24] Nemirovski, A., *Topics in Non-parametric Statistics*, in: M. Emery, A. Nemirovski, D. Voiculescu, Lectures on Probability Theory and Statistics, Ecole d'Eteé de Probabilités de Saint-Flour XXVII – 1998, Ed. P. Bernard. - Lecture Notes in Mathematics **1738**, 87–285.
[25] Nemirovski, A., Shapiro, A., Convex Approximations of Chance Constrained Programs. *SIAM Journal on Optimization* **17:4** (2006), 969-996.
[26] Pinsker M., Optimal filtration of square-integrable signals in Gaussian noise. *Problemy peredachi informatsii*, **16:2** (1980), 120-133. (English transl. in *Problems Inform. Transmission* **16**, 1980.)
[27] Prakasa Rao, B.L.S., *Nonparametric functional estimation.* Academic Press, Orlando, 1983.
[28] Rosenblatt, M., *Stochastic curve estimation.* Institute of Mathematical Statistics, Hayward, California, 1991.
[29] Simonoff, J.S., *Smoothing Methods in Statistics.* Springer, New York, 1996.
[30] Takezawa, K., *Introduction to Nonparametric Regression.* Wiley Series in Probability and Statistics, 2005.
[31] Tsybakov, A.B. *Introduction a l'estimation non-paramétrique.* Springer, 2004.
[32] Wahba, G., *Spline models for observational data.* SIAM, Philadelphia, 1990.
[33] L. Wasserman. *All of Nonparametric Statistics.* Springer Texts in Statistcs. Springer, 2006.

Anatoli Juditsky
Laboratoire Jean Kuntzmann
51 rue des Mathématiques
BP 53
38041 Grenoble Cedex 9 France
E-mail: E-mail: juditsky@imag.fr

Arkadi Nemirovski
School of Industrial and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive
Atlanta GA 30332-0205 USA
E-mail: E-mail: nemirovs@isye.gatech.edu