On Optimum Strategies for Minimizing the Exponential Moments of a Loss Function

Neri Merhav

Department of Electrical Engineering

Technion—Israel Institute of Technology

Haifa 32000, Israel

ISIT 2012, Cambridge, MA, July 2012.

Background and Motivation

Many problems in IT, SP, and related areas are associated with:

$$\min_{s \in \mathcal{S}} \mathbf{E}\ell(X, s),$$

where $X =$ random variable, $s =$ "strategy" (e.g., number, variable, parameter vector, function, etc.), and $\ell(x, s)$ is a loss function.

Examples:

- Compression: $x =$ data, $s =$ code, $\ell(x, s) = -\log s(x) =$ length [bits].
- Estimation: $\boldsymbol{x} = (y, z)$, $s =$ estimator, $\ell(\boldsymbol{x}, s) = [y - s(z)]^2 =$ squared error.
- Quantization: $x =$ data, $s =$ quantizer, $\ell(x, s) = \rho(x - s(x)) =$ error.
- Portfolio selection: $\boldsymbol{x} =$ stock, $\boldsymbol{s} =$ portfolio, $\ell(\boldsymbol{x}, \boldsymbol{s}) = \log(\boldsymbol{s}^T \boldsymbol{x}) =$ wealth.
- Prediction, sequential decision, ...

Background and Motivation (Cont'd)

Minimization of exponential moments

$$\min_{s\in\mathcal{S}} \mathbf{E}e^{\alpha\ell(X,s)} \qquad \alpha > 0$$

received much less attention in IT & SP; more in stochastic control.

Motivations:

- Robustness.
- Risk–sensitivity.
- Related to large deviations performance $\min_s \Pr\{\ell(X,s) \geq L\}$.
- Stronger than $\min_s \mathbf{E}\ell(X,s)$ if minimized by same s for all α.

Q: Can we use knowledge on $\min_s \mathbf{E}\ell(X,s)$ to solve $\min_s \mathbf{E}e^{\alpha\ell(X,s)}$?

Talk Outline

A: Yes, we can!

Outline:

- A simple relationship between the two criteria.
- Some general discussion.
- Several application examples.
- The asymptotic regime.
- Future work (if time permits).

Basic Relationship

Assume $\exists s \in \mathcal{S}$ s.t.

$$Z(s) \stackrel{\Delta}{=} \mathbf{E}_P \exp\{\alpha \ell(X, s)\} < \infty.$$

$s^* \in \mathcal{S}$ minimizes $\mathbf{E}_P \exp\{\alpha \ell(X, s)\}$ if $\exists$ probability distribution Q^* on $\mathcal{X}$ s.t.

1. $s^* = \text{argmin}_{s \in \mathcal{S}} \mathbf{E}_{Q^*} \{\ell(X, s)\}$.
2. $Q^*(x) \propto P(x) e^{\alpha \ell(x, s^*)}$.

Basic Relationship

Assume $\exists s \in \mathcal{S}$ s.t.

$$Z(s) \stackrel{\Delta}{=} \mathbf{E}_P \exp\{\alpha \ell(X, s)\} < \infty.$$

$s^* \in \mathcal{S}$ minimizes $\mathbf{E}_P \exp\{\alpha \ell(X, s)\}$ if $\exists$ probability distribution Q^* on $\mathcal{X}$ s.t.

1. $s^* = \text{argmin}_{s \in \mathcal{S}} \mathbf{E}_{Q^*}\{\ell(X, s)\}$.
2. $Q^*(x) \propto P(x) e^{\alpha \ell(x, s^*)}$.

Proof.

$$\begin{aligned}
\mathbf{E}_P \exp\{\alpha \ell(X, s)\} &= \mathbf{E}_{Q^*} \exp\left\{\alpha \ell(X, s) + \ln \frac{P(X)}{Q^*(X)}\right\} \\
&\geq \exp\left\{\alpha \mathbf{E}_{Q^*} \ell(X, s) - D(Q^* \| P)\right\} \\
&\geq \exp\left\{\alpha \mathbf{E}_{Q^*} \ell(X, s^*) - D(Q^* \| P)\right\} \\
&= \mathbf{E}_P \exp\{\alpha \ell(X, s^*)\}.
\end{aligned}$$

Discussion

- Partially related results: in stochastic control (e.g., Fleming *et al.* '97).
- Related to the Laplace principle: $\ln \mathbf{E}_P e^Y \equiv \sup_Q [\mathbf{E}_Q Y - D(Q\|P)]$.
- Saddle point of $F(s,Q) = \alpha \mathbf{E}_Q \ell(X,s) - D(Q\|P)$.
- $\min_s \max_Q F(s,Q) \;\Leftrightarrow\; \min_s \max_{Q:\, D(Q\|P)\le\epsilon} \mathbf{E}_Q \ell(X,s)$ – robustness.
- Risk–seeking cost:

  $\max_s \mathbf{E} e^{-\alpha\ell(X,s)} \leftrightarrow \min_s \min_Q [\alpha \mathbf{E}_Q \ell(X,s) + D(Q\|P)]$.
- Corresponds to $\min_s \min_{Q:\, D(Q\|P)\le\epsilon} \mathbf{E}_Q \ell(X,s)$.

Applications

Example 1: Lossless Source Coding (Warm-up Exercise)

Here, $\ell(x,s) = -\ln s(x)$, $s =$ probability assignment.

$\mathbf{E}e^{\alpha\ell(X,s)}$ – related to Pr$\{\ell(X,s) \geq L\}$ – buffer overflow.

$$\min_s \mathbf{E}_Q\{-\ln s(X)\} \;\Rightarrow\; s^* = Q.$$

Find

$$Q(x) \propto P(x)e^{-\alpha \ln Q(x)} = \frac{P(x)}{[Q(x)]^\alpha}$$

$$\Rightarrow s^*(x) = Q^*(x) \propto [P(x)]^{1/(1+\alpha)}.$$

Leads to the Rényi entropy

$$H_{1/(1+\alpha)}(P) = \frac{1+\alpha}{\alpha} \ln \left(\sum_{x\in\mathcal{X}} [P(x)]^{1/(1+\alpha)} \right).$$

Example 2: Quantization

$\ell(x,s) = [x - s(x)]^2$, $s : \mathcal{X} \to \{\hat{x}_0, \hat{x}_1, \ldots, x_{M-1}\}$ quantizer.

For $\min_s \mathbf{E}_P[X - s(X)]^2$ – iterative algorithm (Lloyd–Max):

- Given $\hat{x}_0, \hat{x}_1, \ldots, x_{M-1}$, apply NN partioning.
- Given a partition, let $\hat{x}_i =$ centroid of i–th quantization cell.

Consider the risk–seeking cost

$$\max_s \mathbf{E} e^{-\alpha[X-s(X)]^2} \Leftrightarrow \min_s \min_{Q:\, D(Q\|P)\leq\epsilon} \mathbf{E}_Q[X - s(X)]^2.$$

Motivated by friendly pre-processing $P \to Q$: dithering, companding, watermarking... Equivalent to $\min_s \min_Q \{\alpha \mathbf{E}_Q[X - s(X)]^2 + D(Q\|P)\}$. Suggests an iterative Lloyd–Max–like algoritm with two nested loops:

- Inner loop: Given Q, apply Lloyd–Max.
- Outer loop: Given s, calculate $Q(x) \propto P(x) e^{-\alpha[x-s(x)]^2}$.

Example 3: Non–Bayesian Estimation

Let $\boldsymbol{X} \sim \mathcal{N}(\theta \cdot \boldsymbol{u}, \Lambda), \quad \boldsymbol{u} \in \mathbb{R}^n, \ \Lambda \in \mathbb{R}^{n \times n}$.
Estimation: $\ell(\boldsymbol{x}, s) = [\theta - s(\boldsymbol{x})]^2$, $\mathcal{S} =$ all unbiased estimators.
$\mathbf{E}_\theta[\theta - s(\boldsymbol{X})]^2$ minimized by ML estimator:

$$s(\boldsymbol{x}) = \frac{\boldsymbol{u}^T \Lambda^{-1} \boldsymbol{x}}{\boldsymbol{u}^T \Lambda^{-1} \boldsymbol{u}} \triangleq \boldsymbol{v}^T \boldsymbol{x}.$$

Q: What about $\mathbf{E}_\theta e^{\alpha[\theta - s(\boldsymbol{X})]^2}$?
Let us guess that the same s minimizes also the exonentiated square–error.

$$\begin{aligned} Q(\boldsymbol{x}) &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \theta\boldsymbol{u})^T \Lambda^{-1} (\boldsymbol{x} - \theta\boldsymbol{u}) + \alpha \left(\boldsymbol{v}^T \boldsymbol{x} - \theta \right)^2 \right\} \\ &= \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \theta\boldsymbol{u})^T (\Lambda^{-1} - 2\alpha \boldsymbol{v}\boldsymbol{v}^T)(\boldsymbol{x} - \theta\boldsymbol{u}) \right\}, \end{aligned}$$

ML estimator for Q : $\quad s(\boldsymbol{x}) = \dfrac{\boldsymbol{u}^T (\Lambda^{-1} - 2\alpha \boldsymbol{v}\boldsymbol{v}^T)\boldsymbol{x}}{\boldsymbol{u}^T (\Lambda^{-1} - 2\alpha \boldsymbol{v}\boldsymbol{v}^T)\boldsymbol{u}} = \boldsymbol{v}^T \boldsymbol{x}.$

The conditions hold!

Universal Strategies

Consider a situation where $\boldsymbol{X} = (X_1, \ldots, X_n)$ and for every given s, $\mathbf{E}_P e^{\alpha \ell(\boldsymbol{X}, s)}$ is asymptotically exponential in n, that is

$$E(s, \alpha, P) \stackrel{\Delta}{=} \lim_{n \to \infty} \frac{\ln \mathbf{E}_P e^{\alpha \ell(\boldsymbol{X}, s)}}{n} \quad \text{exists.}$$

- s^* is asymptotically optimal: $E(s^*, \alpha, P) \leq E(s, \alpha, P)$ for all $s \in \mathcal{S}$.
- s^* is universal: if in addition s^* is independent of both α and P.

Observation: If $\exists\ s^*$ and a functional $\lambda(Q)$ s.t.

- $\forall\ T_Q, \forall\ \boldsymbol{x} \in T_Q$, $\ell(\boldsymbol{x}, s^*) \leq n[\lambda(Q) + o(1)]$, and
- $\forall\ T_Q, \forall\ s \in \mathcal{S}$, $\left| T_Q \cap \{\boldsymbol{x} :\ \ell(\boldsymbol{x}, s) \geq n[\lambda(Q) - o(1)]\} \right| \geq e^{-no(1)} |T_Q|$,

then s^* is universal and

$$E(s^*, \alpha, P) = \max_Q [\alpha \lambda(Q) - D(Q \| P)].$$

Examples of Universal Strategies

Example 1: Fixed–Rate Lossy Compression: Rate $= R$

- $\boldsymbol{X} \sim P$ is encoded–decoded by s w.r.t. distortion meaure d.
- $\ell(\boldsymbol{x}, s) =$ distortion in reconstructing $\boldsymbol{x}$ using s.
- $D_Q(R) =$ distortion–rate function of Q.

$$\text{Here} \quad \lambda(Q) = D_Q(R).$$

Conditions above met by the covering lemma and its converse.
s^* – based on covering each T_Q by $\sim e^{nR}$ spheres of radius $D_Q(R)$.

Example 2: Variable–Rate Lossy Compression: Distortion $= D$

- $\ell(\boldsymbol{x}, s) =$ description length of $\boldsymbol{x}$ using s.
- $R_Q(D) =$ rate–distortion function of Q.

$$\lambda(Q) = R_Q(D).$$

Related results on guessing with a fidelity criterion [Arikan & M, 1998].

Examples of Universal Strategies (Cont'd)

Example 3: Variable–Rate Lossless Compression

In the lossless case, some refined results are available:
An extension of Rissanen's universal coding theorem [Rissanen '84]:

Given a parametric family of source $\{P_\theta\}$,

$$\frac{1}{\alpha}\ln \mathbf{E}_\theta \exp\{\alpha\ell(\boldsymbol{X},s)\} \geq nH_{1/(1+\alpha)}(P_\theta) + (1-\epsilon)\frac{k}{2}\log n,$$

for every code s and for every θ, except a subset $\mathcal{A}_\epsilon(n)$ whose volume $\to 0$.

For the class of DMS's, there is s^* that satisfied the reversed inequality if $(1-\epsilon)$ is replaced by $(1+\epsilon)$.

Examples of Universal Strategies (Cont'd)

Example 4: Universal Prediction

The universal lossless compression result can be harnessed to obtain a non–trivial lower bound on universal prediction for Gaussian ARMA processes: For example, let $\{X_t\}$ be a Gaussian AR(1) process

$$X_t = \theta \cdot X_{t-1} + W_t, \qquad \{W_t\} \text{ Gaussian i.i.d. with variance } \sigma^2$$

where θ is unknown. Then, denoting $f(x) \stackrel{\Delta}{=} \frac{1}{2}(\frac{1}{x} + \ln x - 1)$:

$$\frac{1}{n} \ln \left[\mathbf{E}_\theta \exp \left\{ \alpha \sum_{t=1}^{n} (X_t - s(X^{t-1})^2 \right\} \right]$$

$$\geq \underbrace{\frac{\alpha\sigma^2}{1-2\alpha\sigma^2} - f(1-2\alpha\sigma^2)}_{\text{bound even for known } \theta} + \underbrace{\frac{(1-\epsilon)\alpha\sigma^2}{1-2\alpha\sigma^2} \cdot \frac{\ln n}{n} + o\left(\frac{\ln n}{n}\right)}_{\text{price of ignorance}}$$

for all $\theta \in (-1, 1)$ except a set $\mathcal{A}_\epsilon(n)$ whose volume $\to 0$.

Thank You!